# Meta-Computing at D0

Igor Terekhov for the D0 collaboration

D0 Run II is one of the two large collider experiments at Fermilab and one of the largest currently running High Energy Physics Experiments in the world. Its amount of data, throughput of data processing, and the size of the collaboration present a unique challenge for the experiment's meta-computing system. To meet the challenge, the SAMGrid system is being developed to allow globally distributed, high-throughput data processing with many Grid features. At the core of the system is the mature data handling system, SAM. We add the Job and Information Management to the data handling to arrive to a complete Grid.

## 1. Introduction

The $\bar{p}p$ collider at Fermilab, the Tevatron, has undergone a major upgrade called Run II; the D0 experiment is one of the two Tevatron collaborations. The experiment is well-positioned to probe the frontiers of the High Energy Physics (HEP) and has challenging research opportunities for the participating physicists [1]. It is no surprise that, as we show below, the computing at the experiment is accordingly challenging, both in terms of the analysis and from the overall system point of view.

Computing per se at a HEP experiment such as D0 is the physics analysis, i.e. the applications. The term *Meta-Computing* can be used to describe the data handling (data grid) and the environment for the applications to run on the Grid. The distinction is somewhat analogous to that between *the data* referring to the actual event data, stored in files on disk or tape, and *the meta-data* referring to the description of the data, stored e.g., in a database. The focus of our paper is the meta-computing at D0. Nevertheless, *computing* in general includes meta-computing, that is why we may use the terms interchangeably in the rest of the paper.

The D0 detector collects, filters and stores the collider event data, which is later processed through reconstruction and additional filtering in order to provide data for analysis. The broad goal of the meta-computing system is to enable such data processing and analysis, by delivering data from producers to storage and from storage to consumers, and by scheduling user jobs accordingly. From the computing point of view, D0 user applications exist solely to process (produce and/or consume) data. Consequently, the data handling challenge is the core of the meta-computing challenge at D0. Here are some of the numbers that may convey the scope of the challenge.

The detector produces data through about a million channels (at least $793 * 10^3$ from the Silicon Microstrip Tracker); about 5-15% of these channels are being read during an event. The digitized event's size as it leaves the Data Acquisition System is about 250KB, to increase by 25% in the second half of the Run (Run IIb). The recorded (after the online filtering) event rate is about 25Hz, which is projected to double in Run IIb. On average, this amounts to about 0.5 Tera-Byte per day of the raw detector data. The data is reconstructed with the same aggregated rate and a similar event size of the output data, thus increasing the total amount of data imported daily into the system to about 1TB.

Whereas it is difficult to predict the exact life-time of the experiment or its aggregate data acquisition rate, in the estimated three years to follow, on the order of $10^9$ events will be collected. Together with the processed types, the grand total dataset set size is expected to be 1-2 Peta-Bytes. In addition, the Monte Carlo simulation data will remain important well into the life-time of the experiment, and the six D0 processing centers will produce about 300 additional TB of data in the next two years.

Impressive as they are, the data amount and processing rates are not the primary factor in the complexity of the computing challenge. There are more than six hundred collaborators in the experiment from about eighty institutions in eighteen countries, which projects as follows onto the complexity of the D0 computing. Many of the participating institutions will host computing resources (and even storage, as in the case with NIKHEF at The Netherlands or IN2P3 in France), as opposed to merely contributing to a centrally located facility. We are therefore observing a dramatic decentralization of the computing resources within D0. One of the main reasons for such decentralization is that the institutions often plan to share the same resources with other Particle Physics experiments. In addition, the D0 collaboration plans to build regional analysis centers in addition to the central analysis site, the FNAL.

Thus, *the degree to which the computing is globally distributed* is an increasingly important factor in the complexity of the D0 computing. From the perspective of the Grid community, the D0 collaboration is a classic example of a large *Virtual Organization*, whose members share resources for solving common problems. Accordingly, the experiment's meta-computing system is a grid system which we describe in the remainder of the paper, starting with the data handling.

## 2. SAM – The Grid-like Data Handling System

In response to the data processing challenge, the D0 experiment together with the FNAL Computing Division started in 1997 a joint project, SAM (Sequential Access using Meta-data), to address the experiment's data handling needs, see the project's Web page at [2]. It's major goals can be summarized as follows:

- Reliably store all the produced data, both detector (real) and Monte-Carlo (simulated), in a Mass Storage System
- Distribute the data globally to analysis centers both within FNAL and beyond
- Catalogue the data contents, *provenance*, status, replica locations, processing history, user datasets etc.
- Manage hardware resources so as to implement the experiment's policies and optimize throughput.

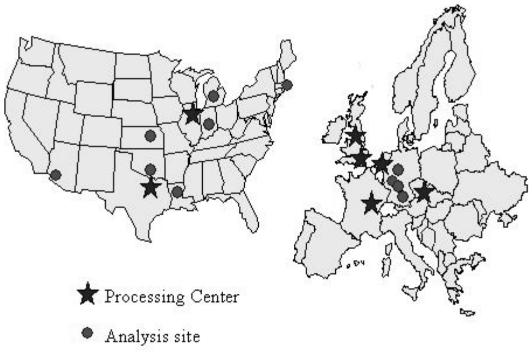

Figure 1. SAM deployment within D0

The history of the SAM project and many of its design decisions have been described in previous papers [3,4] and references therein. Since those reports, the SAM usage has continued to grow, see Fig. 1. In the present paper, we concentrate on the modern perspective onto the system as well as recent developments.

### 2.1. The Storage Systems

Before we proceed to the data handling per se, it is worth highlighting the primary Mass Storage System (MSS) for D0, Enstore [5], where all of the primary, and most of the derived data is permanently stored. It is a grid-enabled, advanced robotic tape-based storage system with a disk cache developed at Fermilab's Integrated System Development department of the Computing Division, in collaboration with the dCache [6] project. From the D0 point of view, the disk cache adds several principal advantages to this storage system, including the ability of the MSS to appear effectively local to the remote installations for efficient export and import of data, see the above references for more information. This is an example of a modern, truly network-capable Storage System that, together with other emerging Storage Systems are changing the data access paradigm.

### 2.2. The Data Handling

The primary role of a modern *Data Handling System* (DHS) is no longer in mediating access to a single Storage System by providing a buffer and assisting in managing its resources [3] by means of e.g. clustering of the data. Rather, the DHS will multiplex the data access by the globally distributed *stations* [3] onto the globally distributed Storage Systems, and thereby provide *Grid collective services* that span multiple facilities (whether storage or processing) [7,8].

The first and foremost service provided is the data replication among the various Storage Systems. In general, the clusters where D0 applications will run will not have access to the storage systems that host the data. Rather, these will have locally configured, highly optimized for local conditions storage systems that will act as cache. The DHS will move data on demand to and from these caches, as a "side effect" of the job scheduling.

Other services include reliability, security and resource management. Reliability at the collective level refers to the ability of the DHS to choose an alternate Storage System when e.g., the data replica from the first choice Storage System fails. Security services consist in translation of the credentials of the user community onto those required by the individual storage systems, and optionally vice versa (for example, the DHS may aggregate multiple requests for the same file from individual users and present its own, "service"-type credential to the MSS, which may vary in the type and issuing authority among the various MSS'). Resource management has been described elsewhere [3].

The SAM project will be affected by the above paradigm shift as follows. Recall that the key concept in the system is the station with its cache. The cache of the station will be viewed as a local Storage System, which is normally intended for applications incapable of network file I/O. The associated cache management will be modularized into the Storage component whereas the rest of the station's functionality together with the global services will comprise the proper Data Handling component, which will be the future focus of the project. At the time of writing this paper, we are seeking to understand the interfaces between the two components so that we can adapt to new Storage Systems.

### 2.3. SAM in the Grid

We started formally to establish a relation between the SAM system and the Grid in our earlier papers [7,8]. Historically, the development of the world-wide Grid community has been happening in the last several years, somewhat parallel to the development of SAM and other D0 systems. The Grid community's focus is on the standards in protocols and interfaces, i.e., in better defining what the (data) Grid is *de jure*. The D0 SAM system, on the other hand, has all the principal functionalities of a *data grid* and can therefore *de facto* be considered such. One of our long-term goals is inter-operability with other experiments and systems. In order for the SAM system to gain the status of a Grid system *de jure*, it will continue to embrace the emerging standards and adopt the maturing grid technologies.

Among other, emerging data grid projects perhaps the most prominent is the European Data Grid project [9], which is broadly expected to deliver a production data grid by the time of the LHC experiments starting to take data. Other noteworthy projects include CrossGrid [10] and NorduGrid [11].

## 3. Towards the Grid

As the experiment's data handling system continues to become a full-fledged grid component, D0 is developing the need for a complete Grid solution, including services for job and information management. By its very definition, Grid work is collaborative in nature and therefore the experiment has been actively engaging in collaboration with other experiments, such as CDF [12], other institutions and other disciplines, most notably Computer Scientists.

The most prominent collaboration is the Particle Physics Data Grid (PPDG) [13]. An example PPDG activity is presented at this conference as well, [14]. The experiment also participates in the GridPP collaboration[16], funded by the governments of the UK and the European Union, as well as in the DutchGrid [17].

Thanks to these fundings and collaboration, it was possible for D0 to spin off of SAM a project for Job and Information Management, JIM, some time this year. Together, SAM and JIM form the SAMGrid project that will handle the expanding needs of the D0 experiment for the Grid computing. Please refer to the presentation [15] at this Conference for a complete technical description of the project, including the architecture, project plans and schedules, as well as the testbed. Below are highlights of the JIM project.

In our model, the experiment's computing will be distributed among a collection of *sites.* Each site will operate one or more *clusters.* Some of those clusters will be fully or partially exposed to and governed by the D0 Grid. A cluster may be a small collection of SMPs or a big farm of Intel/Linux computers. From the experiments Grid point of view, each cluster will provide either some *computing power* or some *data storage capacity* or both. Locally, computational resources as well as Grid jobs will be managed by *Local Resource Management System (LRMS).* LRMS will provide standard minimal but sufficient Grid interface which will be used to:

- provide information about local resource availability
- submit, monitor and control Grid jobs
- store and move data to and from clusters

Users will access the experiment's Grid from their *Grid client* computers. Grid client computer is any computer with Grid client software installed. Initially, we assume a computing model in which the user builds their application interactively on the grid client computer. Users will use Grid to get access to experiment's data and to submit Grid jobs.

The Grid job submission and scheduling is powered by the Condor-G technology [18]. The required key extension of this technology, proposed by the D0 experiment and realized by our collaborators from the Condor [19] team is the promotion of the Condor Match-Making Service, (MMS) [20] to the Grid level. While the MMS is also used by the Resource Broker of the EDG project [9], the pioneering idea of the D0 JIM project is in that the MMS is *the* decision making and brokering entity rather than its base or adviser. Thus, from within the D0 Grid effort we promote interoperability and code reuse and reduce development of own (potentially proprietary) solution to a minimum. This is also an example of the D0 experiment contributing into the development of the core Grid technologies. We begin to collaborate more directly with the EDG project to make our strategies more common.

The D0 monitoring architecture follows the standard Grid Monitoring architecture, see [21]. Our prototype monitoring system was originally inspired by that from the NorduGrid project. As is generally the case with our Grid computing, the principal distinction of our monitoring system is the connection with the data handling system. Specifically, it is possible to navigate from the Grid jobs to their data retrieval projects and further to the display of the data handling system.

In October 2002, the JIM protototype has been officially released. It initially linked three sites, FNAL, Imperial College in London-UK and University of Texas in Arlington, TX. Several more sites are expected to be added by the year end, and the whole system is expected to go in production by April 2002.

## 4. Summary

We have described the meta-computing at D0 today. At the center is the Data Handling system, SAM. We have outlined the present status and the future directions for the SAM project. As D0 and SAM move into the Grid era, we developed the SAMGrid project which includes the JIM effort for job and information management.

## 5. Acknowledgements

We would like to thank everyone at D0 who has contributed to this project, and the many important discussions we have had there. We would like to thank the Fermilab Computing Division for its ongoing support of SAM, especially the ODS, D0CA, and ISD Departments. We acknowledge Vicky White, currently at US DOE for her past contributions into the SAM project, continuing support of the project today and for the many fruitful discussions with her that have shaped our ideas for the future of SAMGrid. We thank Ruth Pordes for her tireless support for our projects and for her executive role in the PPDG collaboration. We acknowledge Rick St Denis and other CDF collaborators for the continuing interest in SAMGrid which has helped CDF adopt the SAM data handling system and thus generate more synergies between the experiments. We thank Frank Wuerthwein of CDF and Igor Mandrichenko of FNAL CD-ISD for the technical discussions on the Grid computing at Run II.

This work is sponsored by DOE contract No. DE-AC02-76CH03000. Additional funding for Grid related SAM work is provided by the DOE SciDAC program through Dzero's association with the Particle Physics Data Grid (PPDG) collaboration. The Imperial College effort is funded through GridPP in the UK.

## REFERENCES

1. The D0 experiment home page, http://www-d0.fnal.gov/
2. The SAM project home page, http://d0db.fnal.gov/sam/
3. V. White *et al.*, "D0 Data handling", plenary talk at The International Symposium on Computing in High Energy Physics (CHEP) 2001, September 2001, Beijing China, in Proceedings. L. Carpenter *et al.*, "SAM Overview and Operational Experience at the Dzero Experiment", ibidem. L. Lueking *et al.*, 'Resource Management in SAM and the D0 Particle Physics Data Grid", ibidem.
4. L. Lueking, et al. "The Data Access Layer for D0 Run II", International Conference on Computing in High Energy Physics 2000, Padova, Italy, January, 2000.
5. The Enstore project home page, http://www-isd.fnal.gov/enstore/
6. The dCache Project, http://www-dcache.desy.de/.
7. V. White *et al.* "SAM and the Particle Physics Data Grid", see [3]
8. I. Terekhov *et al.*, "Distributed Data Access and Resource Management in the D0 SAM System" in Proceedings of 10-th International Symposium on High Performance Distributed Computing (HPDC-10), IEEE Press, July 2001, San-Fransisco, CA
9. P. Kunszt, "Status of the EU DataGrid Project", in these Proceedings. The project home page is at http://www.eu-datagrid.org.
10. M. Kunze, "The CrossGrid Project", in these Proceedings.
11. A. Konstantinov, "The NorduGrid Project: Using Globus Toolkit for Building Grid Infrastructure", in these Proceedings.
12. M. Neubauer, "Computing at CDF", in these Proceedings.
13. The Particle Physics Data Grid home page, http://www.ppdg.net.
14. D. Olson, "Interfacing Interactive Data Analysis Tools with the Grid: The PPDG CS-11 Activity", in these Proceedings.
15. G. Garzoglio, "The SAM-GRID Project: Architecture and Plan", in these Proceedings.
16. The GridPP page, http://www.gridpp.ac.uk/.
17. The DutchGrid home page, http://www.dutchgrid.nl/.
18. J. Frey, T. Tannenbaum, M. Livny, I. Foster, S. Tuecke, "Condor-G: A Computation Management Agent for Multi-institutional Grids", in the same proceedings as [8].
19. The Condor Project home page, http://www.cs.wisc.edu/condor/.
20. R. Raman, M. Livny and M. Solomon, "Matchmaking: Distributed Resource Management for High Throughput Computing", in Proceedings of the Seventh IEEE International Symposium on High Performance Distributed Computing, July 28-31, 1998, Chicago, IL.
21. K. Czajkowski, S. Fitzgerald, I. Foster, C. Kesselman. "Grid Information Services for Distributed Resource Sharing", in the same proceedings as [8].